

【特許請求の範囲】

【請求項 1】

ローカライズされたコンテキストビデオ注釈のための方法であって、
セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割することと、
セグメントの複数の入力モダリティを決定するために、それぞれの前記セグメントを
解析することであって、それぞれの入力モダリティが、前記セグメントのコンテンツの形
態を示す、解析することと、

前記入力モダリティに基づいて、前記セグメントを一組の意味クラスに分類すること
と、

前記一組の意味クラスに基づいて、前記セグメントの注釈を決定することであって、
前記注釈が、前記セグメントと関連付けられた意味コンテキスト情報を示す、決定すること
と、を含む、方法。

10

【請求項 2】

前記セグメントを分類することが、

入力モダリティの分類を決定するために、対応する分類子をそれぞれの前記入力モダ
リティに適用することと、

前記複数の入力モダリティの前記分類に基づいて、前記セグメントの統一分類を決定
することと、

前記統一分類に基づいて、前記セグメントの前記注釈を決定することと、を更に含む
、請求項 1 に記載の方法。

20

【請求項 3】

前記統一分類を決定することが、前記複数の入力モダリティの前記分類を互いに融合
させて、前記統一分類を生成することを更に含む、請求項 2 に記載の方法。

【請求項 4】

前記複数の入力モダリティが、前記セグメントのオーディオ信号から分離されたビデオ
フレームを含み、

前記セグメントを分類することが、ディープビジュアル分類子を前記ビデオフレーム
に適用して、前記セグメントのビジュアル分類を生成することを更に含む、請求項 1 に記
載の方法。

【請求項 5】

前記複数の入力モダリティが、前記セグメントのビデオフレームから分離されたオー
ディオ信号を含み、

前記セグメントを分類することが、

前記オーディオ信号をバックグラウンド信号及び音声信号に分解することと、

オーディオ分類子を前記バックグラウンド信号に適用して、前記セグメントのバック
グラウンドオーディオ分類を生成することと、

感情分類子を前記音声信号に適用して、前記セグメントの感情分類を生成すること
と、を更に含む、請求項 1 に記載の方法。

30

【請求項 6】

前記複数の入力モダリティが、前記セグメントのオーディオビジュアル信号から分離
されたテキスト情報を含み、

前記セグメントを分類することが、

前記セグメントの言語的音声を表す音声テキストを取得することと、

前記テキスト情報を前記音声テキストと整列させることと、

テキストベースの分類子を前記整列させたテキストに適用して、前記セグメントのテ
キスト分類を生成することと、を更に含む、請求項 1 に記載の方法。

40

【請求項 7】

前記ムービーのスクリプトを取得し、更に、前記スクリプトを前記音声テキストと整
列させることを更に含む、請求項 6 に記載の方法。

【請求項 8】

50

前記セグメントを分類することが、
前記複数の入力モダリティのそれぞれの特徴埋め込みを取得することと、
前記特徴埋め込みを組み合わせ、統一埋め込みを生成することと、
意味分類子を前記統一埋め込みに適用して、前記統一分類を決定することと、を更に含む、請求項 1 に記載の方法。

【請求項 9】

前記特徴埋め込みを組み合わせることが、特徴連結を前記特徴埋め込みに適用することを更に含む、請求項 8 に記載の方法。

【請求項 10】

前記注釈が、前記セグメントの前記意味コンテキスト情報を示す一組のキーを含み、それぞれのキーが、値と、強度と、を含み、前記値が、前記セグメントの特徴を示し、前記強度は、前記値が前記セグメントと関連付けられる可能性を示す、請求項 1 に記載の方法。

10

【請求項 11】

前記セグメント化ユニットが、前記ビデオのアクト、シーン、ビート、及びショットのうちの 1 つ以上である、請求項 1 に記載の方法。

【請求項 12】

それぞれの意味クラスが、アクション、危険、ロマンス、フレンドシップ、及びアウトドアアドベンチャーのうちの 1 つに対応する、請求項 1 に記載の方法。

【請求項 13】

コンピュータによって実行されると、前記コンピュータに、ローカライズされたコンテキストビデオ注釈のための方法を実行させる命令を記憶している、非一時的コンピュータ可読記憶媒体であって、前記方法が、

20

セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割することと、
前記セグメントの複数の入力モダリティを生成するために、それぞれのセグメントを解析することであって、それぞれの入力モダリティが、前記セグメントのコンテンツの形態を示す、解析することと、

前記入力モダリティに基づいて、前記セグメントを一組の意味クラスに分類することと、

前記一組の意味クラスに基づいて、前記セグメントの注釈を決定することであって、前記注釈が、前記セグメントと関連付けられた意味コンテキスト情報を示す、決定することと、を含む、非一時的コンピュータ可読記憶媒体。

30

【請求項 14】

前記セグメントを分類することが、
前記入力モダリティの分類を決定するために、対応する分類子をそれぞれの入力モダリティに適用することと、

前記複数の入力モダリティの前記分類に基づいて、前記セグメントの統一分類を決定することと、

前記統一分類に基づいて、前記セグメントの前記注釈を決定することと、を更に含む、請求項 13 に記載のコンピュータ可読記憶媒体。

40

【請求項 15】

前記統一分類を決定することが、前記複数の入力モダリティの前記分類を互いに融合させて、前記統一分類を生成することを更に含む、請求項 14 に記載のコンピュータ可読記憶媒体。

【請求項 16】

前記複数の入力モダリティが、前記セグメントのオーディオ信号から分離されたビデオフレームを含み、

前記セグメントを分類することが、ディープビジュアル分類子を前記ビデオフレームに適用して、前記セグメントのビジュアル分類を生成することを更に含む、請求項 13 に記載のコンピュータ可読記憶媒体。

50

【請求項 17】

前記複数の入力モダリティが、前記セグメントのビデオフレームから分離されたオーディオ信号を含み、

前記セグメントを分類することが、

前記オーディオ信号をバックグラウンド信号及び音声信号に分解することと、

オーディオ分類子を前記バックグラウンド信号に適用して、前記セグメントのバックグラウンドオーディオ分類を生成することと、

感情分類子を前記音声信号に適用して、前記セグメントの感情分類を生成することと、を更に含む、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 18】

前記複数の入力モダリティが、前記セグメントのオーディオビジュアル信号から分離されたテキスト情報を含み、

前記セグメントを分類することが、

前記セグメントの言語的音声を表す音声テキストを取得することと、

前記テキスト情報を前記音声テキストと整列させることと、

テキストベースの分類子を前記整列させたテキストに適用して、前記セグメントのテキスト分類を生成することと、を更に含む、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 19】

前記方法が、前記ムービーのスク립トを取得し、更に、前記スク립トを前記音声テキストと整列させることを更に含む、請求項 18 に記載のコンピュータ可読記憶媒体。

【請求項 20】

前記セグメントを分類することが、

前記複数の入力モダリティの前記分類からそれぞれの特徴埋め込みを取得することと、

前記特徴埋め込みを組み合わせ、統一埋め込みを生成することと、

意味分類子を前記統一埋め込みに適用して、前記統一分類を決定することと、を更に含む、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 21】

前記特徴埋め込みを組み合わせることが、特徴連結を前記特徴埋め込みに適用することを更に含む、請求項 20 に記載のコンピュータ可読記憶媒体。

【請求項 22】

前記注釈が、前記セグメントの前記意味コンテキスト情報を示す一組のキーを含み、それぞれのキーが、値と、強度と、を含み、前記値が、前記セグメントの特徴を示し、前記強度は、前記値が前記セグメントと関連付けられる可能性を示す、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 23】

前記セグメント化ユニットが、前記ビデオのアクト、シーン、ビート、及びショットのうちの 1 つ以上である、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 24】

それぞれの意味クラスが、アクション、危険、ロマンス、フレンドシップ、及びアウトドアアドベンチャーのうちの 1 つに対応する、請求項 13 に記載のコンピュータ可読記憶媒体。

【請求項 25】

ローカライズされたコンテキストビデオ注釈に基づいて広告を配置するための方法であって、

セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割することと、

セグメントの複数の入力モダリティを決定するために、それぞれの前記セグメントを解析することであって、それぞれの入力モダリティが、前記セグメントのコンテンツの形態を示す、解析することと、

10

20

30

40

50

前記入力モダリティに基づいて、前記セグメントを一組の意味クラスに分類することと、

前記一組の意味クラスに基づいて、前記セグメントの注釈を決定することであって、前記注釈が、前記セグメントと関連付けられた意味コンテキスト情報を示す、決定することと、

広告を配置するための標的位置として、前記ビデオファイルのセグメント間のセグメント間可用性 (I S A) を識別することと、

前記 I S A と関連付けられた一組のセグメントの注釈を広告システムに送信することであって、前記一組のセグメントが、前記 I S A の先行セグメント及び前記 I S A の後続セグメントのうちの 1 つ以上を含む、送信することと、を含む、方法。

10

【請求項 26】

ローカライズされたコンテキストビデオ注釈に基づいて自由裁量の視聴を容易にするための方法であって、

セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割することと、

セグメントの複数の入力モダリティを決定するために、それぞれの前記セグメントを解析することであって、それぞれの入力モダリティが、前記セグメントのコンテンツの形態を示す、解析することと、

前記入力モダリティに基づいて、前記セグメントを一組の意味クラスに分類することと、

前記一組の意味クラスに基づいて、前記セグメントの注釈を決定することであって、前記注釈が、前記セグメントと関連付けられた意味コンテキスト情報を示す、決定することと、

20

前記ビデオの視聴者から視聴選好を取得することと、

前記複数のセグメントの前記注釈に基づいて、前記複数のセグメントから一組の視聴セグメントを決定することであって、前記一組の視聴セグメントが、前記視聴選好に従う、決定することと、を含む、方法。

【発明の詳細な説明】

【技術分野】

【0001】

(関連出願の相互参照)

30

本出願は、2019年10月22日に出版され、発明者 Suresh Kumar 及び Raja Bala による「Personalized Viewing Experience and Contextual Advertising in Streaming Media based on Scene Level Annotation of Videos」と題する米国仮出願第 62/924,586 号、代理人整理番号 PARC-20190487US01 の利益を主張するものであり、その開示は、参照により本明細書に組み込まれる。

【0002】

(発明の分野)

本開示は、一般的に人工知能 (artificial intelligence、AI) の分野に関連する。より具体的には、本開示は、ビデオからローカライズされたコンテキスト情報を取得し、ビデオの個々のセグメントに注釈付けするためにローカライズされたコンテキスト情報を利用するためのシステム及び方法に関する。

40

【背景技術】

【0003】

ビデオコンテンツ配信サービス (例えば、YouTube) 及びストリーミングサービス (例えば、Amazon Prime 又は Netflix) の指数関数的成長は、様々なビデオ (例えば、ビデオファイル又はストリーム) への容易なアクセスをもたらしてきた。視聴者は、一般に、特定の年齢層に対する評価及び適合性などの、ビデオと関連付けられたメタデータに関心を持ち得る。ビデオファイルがムービー又はテレビ番組である場合

50

、視聴者はまた、ジャンル、評価、主人公、予告編、及びレビューにも関心を持ち得る。しかしながら、そのような情報は、ビデオの全体的な見解を示すことができ、また、ビデオがシーンからシーンへ進むにつれて、一時的な情報又はガイダンスを提供する。

【0004】

更に、ビデオは、広告サポート形態で視聴者に提供され得る。しかしながら、ビデオに挿入される広告には、ビデオの先行セグメント（すなわち、広告の前に放送していたセグメント）に対していずれのコンテキストも有しない場合がある。ビデオセグメントは、視聴者の感情状態に影響を与え得るので、広告のコンテンツは、ビデオセグメントと関連付けられたコンテキスト情報を考慮することがなければ、視聴者の感情状態に好適ではない場合がある。その結果、広告の配置が逆効果になり得る。

10

【0005】

グローバルなメタデータは、数多くの望ましい特徴をビデオコンテンツプロバイダにもたらすが、効率的な視聴及びマネタイゼーションに対するビデオのローカライズされたコンテキスト注釈を容易にする際に、いくつかの課題が未解決のままである。

【発明の概要】

【0006】

本明細書に記載される実施形態は、ローカライズされたコンテキストビデオ注釈のためのシステムを提供する。操作中に、システムは、セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割し、セグメントの複数の入力モダリティを生成するためのそれぞれのセグメントを解析することができる。それぞれの入力モダリティは、セグメントのコンテンツの形態を示すことができる。次いで、システムは、入力モダリティに基づいて、セグメントを一組の意味クラスに分類し、一組の意味クラスに基づいて、セグメントのための注釈を決定することができる。注釈は、セグメントの意味コンテキスト情報を示すことができる。

20

【0007】

この実施形態の変形例では、システムは、入力モダリティの分類を決定するために、対応する分類子をそれぞれの入力モダリティに適用することができる。続いて、システムは、複数の入力モダリティの分類に基づいて、セグメントの統一分類を決定することができる。次いで、システムは、統一分類に基づいて、セグメントの注釈を決定することができる。

30

【0008】

更なる変形例では、システムは、複数の入力モダリティの分類を互いに融合させて統一分類を生成することによって、統一分類を決定することができる。

【0009】

この実施形態の変形例では、複数の入力モダリティは、セグメントのオーディオ信号から分離されたビデオフレームを含む。次いで、システムは、ディープビジュアル分類子をビデオフレームに適用してセグメントのビジュアル分類を生成することによって、セグメントを分類することができる。

【0010】

この実施形態の変形例では、複数の入力モダリティは、セグメントのビデオフレームから分離されたオーディオ信号を含む。次いで、システムは、オーディオ信号をバックグラウンド信号及び音声信号に分解することによって、セグメントを分類することができる。次いで、システムは、オーディオ分類子をバックグラウンド信号に適用して、セグメントのバックグラウンドオーディオ分類を生成することができる。システムはまた、感情分類子を音声信号に適用して、セグメントの感情分類を生成することもできる。

40

【0011】

この実施形態の変形例では、複数の入力モダリティは、セグメントのオーディオビジュアル信号から分離されたテキスト情報を含む。次いで、システムは、セグメントの言語的音声を表す音声テキストを取得することによって、セグメントを分類することができる。次いで、システムは、テキスト情報を音声テキストと整列させ、整列させたテキストに

50

テキストベースの分類子を適用して、セグメントのテキスト分類を生成することができる。

【0012】

更なる変形例では、システムは、ムービーのスクリプトを取得し、更に、スクリプトを音声テキストと整列させることができる。

【0013】

この実施形態の変形例では、システムは、複数の入力モダリティの分類からそれぞれの特徴埋め込みを取得し、特徴埋め込みを組み合わせ、統一埋め込みを生成することによって、セグメントを分類することができる。次いで、システムは、意味分類子を統一埋め込みに適用して、統一分類を決定することができる。

10

【0014】

更なる変形例では、システムは、特徴連結を特徴埋め込みに適用することによって、特徴埋め込みを組み合わせることができる。

【0015】

この実施形態の変形例では、注釈は、セグメントの意味コンテキスト情報を示す一組のキーを含むことができる。それぞれのキーは、値及び強度を含むことができる。値は、セグメントの特徴を示すことができ、強度は、値がセグメントと関連付けられる可能性を示すことができる。

【0016】

この実施形態の変形例では、セグメント化ユニットは、ビデオのアクト、シーン、ビート、及びショットのうちの1つ以上であり得る。

20

【0017】

この実施形態の変形例では、それぞれの意味クラスは、アクション、危険、ロマンス、フレンドシップ、及びアウトドアアドベンチャーのうちの1つに対応し得る。

【0018】

本明細書に記載される実施形態は、ローカライズされたコンテキストビデオ注釈に基づいて広告を配置するためのシステムを提供する。操作中に、システムは、セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割し、セグメントの複数の入力モダリティを生成するためのそれぞれのセグメントを解析することができる。それぞれの入力モダリティは、セグメントのコンテンツの形態を示すことができる。次いで、システムは、入力モダリティに基づいて、セグメントを一組の意味クラスに分類し、一組の意味クラスに基づいて、セグメントのための注釈を決定することができる。続いて、システムは、広告を配置するための標的位置として、ビデオファイルのセグメント間のセグメント間可用性(inter-segment availability、ISA)を識別し、ISAと関連付けられた一組のセグメントの注釈を広告システムに送信することができる。一組のセグメントは、ISAの先行セグメント及びISAの後続セグメントのうちの1つ以上を含む。

30

【0019】

本明細書に記載される実施形態は、ローカライズされたコンテキストビデオ注釈に基づいた自由裁量の視聴を容易にするためのシステムを提供する。操作中に、システムは、セグメント化ユニットに基づいて、ビデオを複数のセグメントに分割し、セグメントの複数の入力モダリティを生成するためのそれぞれのセグメントを解析することができる。それぞれの入力モダリティは、セグメントのコンテンツの形態を示すことができる。次いで、システムは、入力モダリティに基づいて、セグメントを一組の意味クラスに分類し、一組の意味クラスに基づいて、セグメントのための注釈を決定することができる。続いて、システムは、ビデオの視聴者から視聴選好を取得し、複数のセグメントの注釈に基づいて、複数のセグメントから一組の視聴セグメントを決定することができる。一組の視聴セグメントは、視聴選好に従う。

40

【図面の簡単な説明】

【0020】

【図1A】本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈を

50

容易にする例示的なビデオ注釈システムを示す。

【0021】

【図1B】本出願の一実施形態による、例示的なビデオのローカライズされたコンテキスト注釈を示す。

【0022】

【図2A】本出願の一実施形態による、ビデオセグメントのマルチモーダルコンテキスト情報からの例示的な注釈生成を示す。

【0023】

【図2B】本出願の一実施形態による、ビデオセグメントと関連付けられた例示的なローカライズされた注釈を示す。

【0024】

【図3A】本出願の一実施形態による、分類ベースのマルチモーダル融合を使用して、ビデオセグメントのローカライズされたコンテキスト注釈を生成する例示的なプロセスを示す。

【0025】

【図3B】本出願の一実施形態による、埋め込みベースのマルチモーダル融合を使用して、ビデオセグメントのローカライズされたコンテキスト注釈を生成する例示的なプロセスを示す。

【0026】

【図4】本出願の一実施形態による、ローカライズされたコンテキスト情報に基づく、ビデオの例示的な広告配置プロセスを示す。

【0027】

【図5A】本出願の一実施形態による、分類ベースのマルチモーダル融合を使用して、ビデオセグメントのマルチモーダル分類を生成する、ビデオ注釈システムの方法を示すフローチャートを示す。

【0028】

【図5B】本出願の一実施形態による、埋め込みベースのマルチモーダル融合を使用して、ビデオセグメントの統一分類を生成する、ビデオ注釈システムの方法を示すフローチャートを示す。

【0029】

【図5C】本出願の一実施形態による、ビデオセグメントの統一分類に基づいて注釈を生成する、ビデオ注釈システムの方法を示すフローチャートを示す。

【0030】

【図6】本出願の一実施形態による、広告をビデオに配置するための注釈情報を提供する、ビデオ注釈システムの方法を示すフローチャートを示す。

【0031】

【図7】本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈をサポートする、ビデオ注釈システムを容易にする例示的なコンピュータシステムを示す。

【0032】

【図8】本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈を容易にする、例示的な装置を示す。

【0033】

図面中、同じ参照番号は、同じ図形要素を指す。

【発明を実施するための形態】

【0034】

以下の説明は、当業者が実施形態を製造及び使用することを可能にするために提示され、特定の用途及びその要件に関連して提供される。開示される実施形態に対する様々な修正は、当業者には容易に明らかとなり、本明細書に定義される一般原理は、本開示の趣旨及び範囲から逸脱することなく、他の実施形態及び用途に適用され得る。したがって、本明細書に記載される実施形態は、示される実施形態に限定されるものではなく、本明細

10

20

30

40

50

書に開示される原理及び特徴と一致する最も広い範囲を与えられるものである。

概要

【0035】

本明細書に記載される実施形態は、(i) ビデオを有意なセグメントに分割し、それぞれのビデオセグメントのコンテンツを示す複数の入力モダリティを取得することによって、(ii) 対応する形態の人工知能(AI)モデルを入力それぞれのモードに適用することによってマルチモーダルコンテキスト情報を測定することによって、及び(iii) マルチモーダルコンテキスト情報を統一して、統一されたコンテキスト情報を取得することによって、ビデオのそれぞれのビデオセグメントからローカライズされたコンテキスト情報を生成するという問題を解決する。

10

【0036】

既存の技術の場合、いくつかのビデオ提供サービスは、しばしば、ビデオの検索及び読み出しのためのシーンレベルコンテンツの検出、要約、サブタイトルの位置合わせ、及びアクター識別情報を使用する。しかしながら、検出されたコンテンツは、その特定のシーンが子供の前で視聴するのに適しているか、又はシーン内に表される感情の種類に適しているかなどの、シーンのコンテキスト情報を表さない場合がある。その結果、視聴者がビデオ内の特定の種類のシーン(例えば、R指定のアクティビティ又は激しい戦闘シーン)の存在に気付く場合であっても、視聴者は、そのビデオを見る前にそのシーンの位置に気付かない場合がある。更に、一部のビデオプロバイダは、広告をビデオの特定の場所に挿入する場合がある。しかしながら、広告の前に再生されたビデオのコンテンツは、視聴者を、広告に適合しない又は好適ではない場合がある感情状態にする場合がある。

20

【0037】

この問題を解決するために、ビデオ注釈システムは、1つ以上のシーンセグメント化技術を使用して、ビデオの個々のセグメントを識別することができる。システムは、各セグメントの開始及び終了タイムスタンプを決定することができる。セグメント化技術は、オブジェクト検出、集合論、及びディープニューラルネットワークのうちの1つ以上に基づくことができる。ビデオのセグメント(又はビデオセグメント)は、ビデオのアクト、シーン、ビート、又はショットに基づくことができる。次いで、システムは、それぞれのセグメントと関連付けられたコンテキスト意味情報を取得することができる。続いて、システムは、セグメントを分類し、分類に基づいてセグメントに注釈付けすることができる。例えば、システムは、セグメントと関連付けられた1つ以上のカテゴリによって決定することができる。したがって、システムは、それぞれの決定されたカテゴリ及びセグメントがそのカテゴリ内にある可能性をセグメントに注釈付けすることができる。次いで、注釈を使用して、ムービーを特定の観衆向きにパーソナライズすることができる(例えば、未成年者が存在する場合に、不適切なビデオセグメントをスキップすることができる)。

30

【0038】

いくつかの実施形態では、システムは、複数の入力モダリティを決定することができ、これは、ビデオと関連付けられたマルチモーダル入力と称することもできる。そのような入力モダリティは、ビデオチャンネル、ビデオのオーディオ信号(例えば、バックグラウンド及び音声)、ビデオのテキスト情報(例えば、スクリプト及びサブタイトル)、並びにビデオの追加的なメタデータ(例えば、ビデオに埋め込まれた既知のメタデータ及び追加的な情報)を含むことができるが、これらに限定されない。システムは、対応する形態のAIモデルを入力各モードに適用し、意味カテゴリに対応する分類を決定することができる。例えば、システムは、ディープビジュアル分類子をビデオチャンネルに適用して、ビデオセグメントの意味カテゴリを識別する(又は推定する)ことができる。同様に、システムは、自然言語処理(natural language processing、NLP)をセグメントのスクリプトに適用して、意味カテゴリも推定することができる。このようにして、システムは、(すなわち、分類の複数のモードから)セグメントのマルチモーダル分類を取得し、同時推論をマルチモーダル入力からの分類に適用することによって統一分類を決定することができる。

40

50

【0039】

次いで、システムは、それぞれのセグメントの注釈を決定し、注釈をビデオに組み込むことができる。注釈は、一組のキーの形態とすることができる。各キーは、セグメントと関連付けられたオブジェクト、ジャンル、コンセプト、又はアクションを示すことができる。キーはまた、評価、キャラクタの名称、及び位置を示すこともできる。各キーは、値及び強度によって表すことができる。それ故に、キーは、{キー：(値、強度)}の形態とすることができる。キーの値は、キーのタイプに依存することができる。例えば、キーがセグメントのカテゴリ（又はジャンル）である場合、値は、アクション、アドベンチャー、コメディ、クライム、ドラマ、ファンタジー、ホラー、ミュージカル、ミステリー、ロマンス、及びSFのうちの1つであり得る。強度は、所与の値（例えば、強度レベル又は信頼レベル）が存在する可能性を示す数字（例えば、0～1）であり得る。 10

【0040】

キーが意味カテゴリを表す場合、システムは、キーと関連付けられた意味カテゴリ（すなわち、値）の正規化された表現を生成することができる。更に、システムは、各キーと関連付けられた値の強度の正規化された表現を示すことができる。正規化は、1（又は100）の値に基づくことができる。セグメントが特定のカテゴリにある可能性は、0～1の小数値で表され、全てのカテゴリの小数値の和は1であり得る。特定のカテゴリと関連付けられた小数值は、注釈値と称することもできる。いくつかの実施形態では、セグメントと関連付けられた各注釈値は、独特な表現（例えば、色、フィルパターン、又は値を有する）で表して、注釈インジケータを生成することができる。続いて、システムは、注釈インジケータをビデオプロバイダに提供することができ、次いで、対応するビデオセグメントを有する注釈インジケータを示すことができる。 20

【0041】

注釈は、視聴者のパーソナライズされた視聴オプションを容易にすることができる。視聴者は、特定のカテゴリに属するビデオのアイテムを自動的にスキップするように選択し得る。カテゴリを示す値の強度が閾値よりも大きい場合、システムは、セグメントがそのカテゴリに属すると判定し得る。それ故に、各セグメントは、複数のカテゴリに属し得る。視聴者はまた、特定の特徴にマッチするビデオの特定の部品を見ることも選択し得る。例えば、視聴者は、特定のアクターを有する、又は特定のアクションを有するセグメントのみを見ることを選択することができる。システムは、視聴者が、注釈の特定のキーの一組の値（例えば、一組のカテゴリ）を選択することを可能にし得る。次いで、システムは、カテゴリに対応するセグメントをスキップするか、そののみを示すことができる。システムは、それぞれのセグメントの開始及び終了タイムスタンプを維持することができる。その結果、選択されたセグメントのタイムスタンプが決定されるたびに、本システムは、開始及び終了タイムスタンプの間のセグメントをスキップするか、示すことができる。 30

【0042】

更に、ビデオのコンテキスト的に関連する広告を配置するために、注釈を使用することができる。システムは、ビデオのセグメント間可用性（inter-segment availability、ISA）（すなわち、セグメント間の利用できるスポット）を広告スポットとして識別することができる。ISAは、シーン間とすることができるので、広告のISAベースの配置は、ビデオ内の重要なシーケンス又はシーンを中断し得ない。その結果、広告の配置は、視聴者に生じさせるいらだちをより少なくし得る。アプリケーションを提供するビデオの管理者は、1つ以上の広告を示すためにビデオを停止させることができる回数を構成し得る。この数に基づいて、システムは、1つ以上のポリシーを使用して広告を配置するための一組のISAを推奨し得る。例えば、ポリシーは、手に汗を握る展開を回避する又は使用すること、シーンの断片化を回避すること、及び心地よい先行セグメントを含むことができる。 40

【0043】

広告スポットが接近したときに、システムは、注釈を広告システム（例えば、広告交換及び／又はプラットフォーム）に提供することができる。注釈は、広告スポット（すな 50

わち、I S A) の先行及び後続のシーン、先行シーンの最後のビート、及び後続シーンの最初のビートと関連付けることができる。広告システムは、注釈と、利用可能な広告のコンテンツとを動的にマッチさせ、注釈に表されるコンテキスト情報にマッチする広告を配置することができる。代替的に、広告の所有者（例えば、保険会社）は、事業に関連するタグ（例えば、火事、爆発など）を購入することができる。注釈が特定のタグにマッチする場合、対応するスポットに広告を配置することができる。このマッチは、キーの値及び対応する強度に基づくことができる。このようにして、注釈は、ビデオのコンテキスト的に関連する広告の配置を容易にすることができる。

例示的なシステム

【0044】

図1Aは、本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈を容易にする例示的なビデオ注釈システムを示す。ビデオ管理環境100は、アプリケーションサーバ104を含み、これは、ビデオコンテンツをホストすることができ、かつビデオ140をユーザ150に提供し得る1つ以上のアプリケーションをホストすることができる。例えば、アプリケーションは、ビデオ140をユーザ150のユーザデバイス106を提供するために、ビデオストリーミング又はテレビベースのコンテンツデリバリサービスを提供することができる。そのようなアプリケーションは、ビデオコンテンツ配送サービス、ビデオストリーミングサービス、テレビジョンサービス、及びバーチャル／拡張現実サービスプロバイダを含むが、これらに限定されない。アプリケーションサーバ104は、広告をビデオ140に組み込むことができる広告プロバイダと通信することができる。アプリケーションサーバ104は、ビデオ140などのビデオコンテンツ及び／又はビデオ140の広告コンテンツを記憶するためのデータベース182を維持することができる。アプリケーションサーバ104は、ローカル又はワイドエリアネットワークであり得るネットワーク130を介して（例えば、デバイス106から）到達可能であり得る。

【0045】

既存の技術については、アプリケーションサーバ104上のアプリケーションは、ビデオ140のシーン142上のシーンレベルコンテンツ検出を使用し得る。シーンレベルコンテンツ検出を使用することで、アプリケーションは、ユーザデバイス106のディスプレイデバイス108からのビデオ検索及び取り出しを容易にすることができる。更に、アプリケーションはまた、シーン142と関連付けられた要約、サブタイトルの位置合わせ、及びアクター識別情報をデバイス106に提供することもできる。しかしながら、シーン142の検出されたコンテンツは、シーン142が子供の前で視聴するのに適しているか、又はシーン142内に表される感情の種類に適しているかなどの、シーン142のコンテキスト情報を表さない場合がある。

【0046】

その結果、ユーザ140がビデオ140に現れ得る特定の種類のコンテンツ（例えば、R指定のアクティビティ又は激しい戦闘シーン）の存在に気付く場合であっても、ユーザ150は、シーン142などの特定のシーンのそのようなコンテンツに気付かない場合がある。更に、アプリケーションは、シーン142の後に、広告をビデオ140に挿入する場合がある。しかしながら、広告を表示する前に再生されたシーン142のコンテンツは、ユーザ150を、広告に適合しない又は好適ではない場合がある感情状態にする場合がある。

【0047】

この問題を解決するために、環境100は、ビデオ注釈システム110を含むことができ、これは、セグメントレベルのコンテキスト情報を決定し、その情報を使用して、ビデオ140のそれぞれのセグメントに注釈付けすることができる。システム110は、セグメント化モジュール112、分析モジュール114、及び注釈モジュール116を備えることができる。システム110は、随意に、広告モジュール118（破線で表す）を含み得る。セグメント化モジュール112は、ビデオ140のセグメント146及び148

などの個々のセグメントを識別することができる、1つ以上のシーンセグメント化技術をサポート及び実装することができる。セグメントがシーンレベルで決定される場合、セグメント146は、シーン142と同じであり得る。

【0048】

操作中に、セグメント化モジュール112は、ビデオ140の各セグメントの開始及び終了タイムスタンプを決定することができる。セグメント化技術は、オブジェクト検出、集合論、及びディープニューラルネットワークのうちの1つ以上に基づくことができる。ビデオセグメントと称することもできるビデオ140のセグメントは、ビデオ140のアクト、シーン、ビート、又はショットであり得る。次いで、分析モジュール114は、セグメント146と関連付けられたコンテキスト意味情報を取得することができる。続いて、分析モジュール114は、セグメント146を分類し、その分類に基づいてセグメント146に注釈付けすることができる。ここで、分析モジュール114は、セグメント146と関連付けられた1つ以上のカテゴリを決定することができる。

10

【0049】

いくつかの実施形態では、分析モジュール114は、ビデオ140と関連付けられた複数の入力モダリティを決定することができる。そのような入力モードは、ビデオチャンネル又はビデオ140のフレーム、ビデオ140のオーディオ信号、及びビデオ140のテキスト情報を含むことができるが、これらに限定されない。オーディオ信号は、バックグラウンドサウンド及び音声の両方を含むことができ、テキスト情報は、ビデオ140のSCRIPT及びビデオ140のサブタイトル（又は字幕）を含むことができる。セグメント化モジュール112は、それぞれの入力モードをセグメント化して、ビデオ140のビデオセグメントと整列させ、分析モジュール114のマルチモーダル入力セット120を生成することができる。例えば、セグメント化モジュール112は、セグメント146（例えば、シーン142）に属するビデオ140のビデオフレーム及びオーディオ信号をセグメント化して、それぞれ、ビデオフレーム122及びオーディオ信号124を決定することができる。同様に、セグメント化モジュール112は、ビデオ140のテキスト情報をセグメント化して、セグメント146と関連付けられたテキスト情報126を決定することができる。入力セット120はまた、セグメント146と関連付けられた追加的なメタデータ情報128（例えば、ビデオ140の既知のメタデータ及び追加的な埋め込み情報）を含むこともできる。このようにして、入力セット120は、複数の入力モダリティ又は入力モードを含むことができる。

20

30

【0050】

分析モジュール114は、対応する形態のAIモデルを入力セット120の各入力モードに適用し、意味カテゴリへの対応する分類を示す意味コンテキスト情報を推定することができる。例えば、分析モジュール114は、ディープビジュアル分類子をビデオフレーム122に適用して、セグメント146のビデオ分類を推定することができる。分析モジュール114はまた、オーディオ分類子及び／又は音声認識をオーディオ信号124に適用して、セグメント146の音声分類を推定することができる。同様に、分析モジュール114は、NLPベースの分類子をテキスト情報126に適用して、セグメント146のテキスト分類を推定することができる。このようにして、分析モジュール114は、（すなわち、複数の分類モードから）セグメント146のマルチモーダル分類を取得することができる。次いで、分析モジュール114は、併せて、セグメント146のマルチモーダル分類（又は分類から取得されるそれぞれの埋め込み）から、融合又は統一分類を同時推論することができる。

40

【0051】

統一分類に基づいて、注釈モジュール116は、注釈インジケータ144を生成して、それぞれの決定されたカテゴリ及びセグメント146がそのカテゴリにある可能性をセグメント146に注釈付けすることができる。注釈モジュール116は、セグメント146（例えば、シーン142）を示しながら、注釈インジケータ144をデータ構造内のセグメント146にマッピングし、及び／又は注釈インジケータ144をディスプレイデバ

50

イス108上に示し得る。次いで、ユーザ150は、注釈インジケータ144を使用して、セグメント146を見るかどうかを判定することができる。例えば、ユーザ150は、未成年者が存在する場合に、セグメント146をスキップし得る。

【0052】

更に、広告モジュール118は、コンテキスト的に関連する広告をビデオ140に配置するために、コンテキスト情報を使用することができる。広告モジュール118は、ビデオ140内のISAを潜在的広告スポットとして識別することができる。ISAは、シーン間とすることができるので、広告のISAベースの配置は、ビデオ140内の重要なシーケンス又はシーンを中断し得ない。その結果、広告の配置は、ユーザ150に生じさせるいらだちをより少なくし得る。アプリケーションサーバ104上で動作しているアプリケーションの管理者は、1つ以上の広告を示すためにビデオ140を停止することができる回数を構成することができる。この数に基づいて、広告モジュール118は、1つ以上のポリシーを使用して広告を配置するための一組のISAを推奨し得る。例えば、ポリシーは、手に汗を握る展開を回避する又は使用すること、シーンの断片化を回避すること、及び対応する注釈によって示される心地よい先行セグメントを含むことができる。

【0053】

ビデオ140のセグメント146と148との間のISAが広告スポットとして選択されたと仮定する。ビデオ140が動作中である間に広告スポットに接近してくると、広告モジュール118は、関連付けられたコンテキスト情報を広告プラットフォームに提供することができる。コンテキスト情報は、セグメント146及び148、セグメント146内の最後のビート、並びにセグメント148内の最初のビートと関連付けることができる。コンテキスト情報は、セグメント146及び148、セグメント146内の最後のビート、並びにセグメント148内の最初のビートの各々を表す1つ以上のキーワードを含むことができる。広告モジュール118は、コンテキスト情報を広告システム（例えば、広告交換及び／又はプラットフォーム）に提供することができ、これは、情報共有プロトコルに基づいてシステム110と通信することができる。プロトコルは、供給側及び／又は需要側の広告配置を容易にすることができる。

【0054】

広告システムは、コンテキスト情報と利用可能な広告のコンテンツとを動的にマッチさせ、コンテキスト情報にマッチする広告を配置することができる。例えば、広告システムが広告交換である場合、広告システムは、コンテキスト情報を入札者に提供することができる。次いで、コンテキスト情報（例えば、キーワード）を使用して、ISA内の正しい広告にマッチさせることができる。代替的に、広告の所有者は、事業に関連するタグを購入することができる。コンテキスト情報が特定のタグにマッチする場合、広告は、セグメント146と148との間の対応するスポットに配置されることができる。このようにして、広告モジュール118は、ビデオ140のコンテキスト的に関連する広告の配置を容易にすることができる。

セグメントレベルの注釈

【0055】

図1Bは、本出願の一実施形態による、例示的なビデオのローカライズされたコンテキスト注釈を示す。ビデオ140は、アクト152を含む一組のアクトを含むことができる。アクトは、ビデオ140のストーリーアークを表すことができる。それぞれのアクトは、いくつかのシーンを含むことができる。シーンは、タイムフレーム及び／又は位置に関して制限されるアクションのユニットとすることができる。ビデオ140のアクションの時間又は位置が移動した場合、一般に、新しいシーンとみなされる。アクト152は、シーン142を含む一組のシーンを含むことができる。それぞれのシーンは、更にビートに分解することができる。この実施例では、シーン142は、ビート154を含むことができる一組のビートを含むことができる。それぞれのビートは、シーンと関連付けられたナレーションを表すことができる。それぞれのビートは、1つ以上のショットを含むことができる。各ショットは、ナレーションの一部を含むことができる。ここで、ビート1

10

20

30

40

50

54は、ショット156を含む一組のショットを含むことができる。例えば、シーン142が教室に対応する場合、シーン142のビートは、指示を提供する教師を示すことができる。そのビートのショットは、指示の送達プロセス中に、学生の表情を示すことができる。

【0056】

セグメント化ユニットが特定の階層要素に設定されたとき、セグメント化モジュール112は、ビデオ140のセグメント化を容易にすることができる。セグメント化モジュール112は、各セグメントがビデオ140の意味部分を表し、セグメント内でコンテキストに関連するような方法でビデオ140を分解し得る。したがって、それぞれのセグメントは、ストーリーの意味論的に有意なピースを提供することによって、ビデオ140のストーリーを進め得る。例えば、ビデオ140を分解するための意味コンテキスト方法がシーンによるものである場合、セグメント化モジュール112は、シーンに基づいて、ビデオ140をセグメント化することができる。したがって、ビデオ140のセグメント146は、シーン142を表すことができる。したがって、ビデオ140のそれぞれのセグメントは、シーンに対応することができる。セグメント化モジュール112は、シーン142を、シーン142内の物語ビートレベルで更に分解することができる。

【0057】

分析モジュール114は、それぞれのセグメントと関連付けられたマルチモーダルコンテキスト情報を決定することができる。マルチモーダルコンテキスト情報は、セグメントの結合干渉から取得されるマルチモーダル分類を含むことができる。分析モジュール114は、随意に、セグメントの内のサブセグメントと関連付けられたコンテキスト情報を決定することができる。続いて、注釈モジュール116は、各セグメントに、対応するコンテキスト情報を注釈付けすることができる。例えば、分析モジュール114は、セグメント146（シーン142）と関連付けられたマルチモーダルコンテキスト情報170を決定することができる。次いで、注釈モジュール116は、セグメント146にコンテキスト情報170を注釈付けすることができる。

【0058】

セグメント146に、対応するコンテキスト情報170を注釈付けするために、注釈モジュール116は、セグメント146のセグメント識別子172とコンテキスト情報170との間のマッピングを維持し得る。セグメント識別子172は、ビデオ140のビデオ識別子及びシーン142のシーン番号のうちの1つ以上を含むことができる。ビデオプロバイダアプリケーションは、そのデータベース（例えば、図1Aのデータベース182）内のマッピングを維持することができる。コンテキスト情報170は、マルチモーダル分類（又は分類から取得されるそれぞれの埋め込み）を融合させる又は統一することによって生成される融合分類を含むことができる。

【0059】

いくつかの実施形態では、コンテキスト情報170は、ビートレベルでコンテキスト情報を含むことができる。シーンレベルコンテキスト情報170は、必ずしも、対応するビートレベルコンテキスト情報の集合体又は複合体でない場合がある。例えば、シーン142（セグメント146）は、ビート154などのいくつかのビートに更に分解されることができる。分析モジュール114は、シーン142内のそれぞれのビートと関連付けられたマルチモーダルコンテキスト情報を決定することができる。次いで、コンテキスト情報170は、シーンレベル情報、及びビート154と関連付けられたビートレベル情報の両方を含むことができる。これは、システム110が、セグメント146と関連付けられたコンテキスト情報のより細かい分析を利用することを可能にする。

【0060】

更に、ビジュアルキューをスクリーン上に提供するために、注釈モジュール116は、コンテキスト情報170から注釈インジケータ144と関連付けられた情報を生成し、生成された情報をセグメント146の識別子と関連付けて記憶し得る。次いで、システム110は、注釈インジケータ144をビデオプロバイダサービス（例えば、図1Aのアプリ

10

20

30

40

50

リケーションサーバ104上で動作しているアプリケーション)に提供することができる。ビデオプロバイダサービスは、注釈インジケータ144にシーン142を表示して、コンテキスト情報170のビジュアルキューを視聴者に提供し得る。

【0061】

図2Aは、本出願の一実施形態による、ビデオセグメントのマルチモーダルコンテキスト情報からの例示的な注釈生成を示す。セグメント120の入力セット120は、セグメント146と関連付けられたビデオフレーム122、オーディオ信号124、及びテキスト情報126を含むことができる。ビデオ140がムービーである場合、ビデオ140は、スクリプトを伴い得る。典型的に、スクリプトは、一連のシーンとして十分に構造化することができる。各シーンに対して、スクリプトは、位置及び時間を記載しているシーン番号及び見出しを含むことができる。例えば、ビデオ140が、第二次世界大戦に基づき得る「ムービーX」というタイトルのムービーである場合、セグメント146/シーン142は、ムービーXのあるシーンであり得る。次いで、システム110は、パブリックにアクセス可能なリソース(例えば、ムービースクリプトを記憶するためのウェブリポジトリ)からムービーXのスクリプトを取得することができる。

【0062】

スクリプトをセグメント化することによって、次いで、セグメント化モジュール112は、シーン142と関連付けられたムービーXのスクリプトのスクリプトセグメント204を含み得るテキスト情報126を生成することができる。スクリプトセグメント204は、シーン142がシーン77の始まりであることを示すことができる。スクリプトセグメント204はまた、シーン142が第二次世界大戦の西部戦線の外部の位置にあると述べている見出しを含むこともできる。したがって、セグメント識別子172は、シーン番号77を組み込むことができる。セグメント化モジュール112はまた、ビデオ140をセグメント化することによって、スクリプトセグメント204に対応するビデオセグメント202を取得することもできる。ビデオセグメント202から、セグメント化モジュール112は、ビデオフレーム及び対応するオーディオ信号を分離して、それぞれ、ビデオフレーム(又はビデオチャンネル)122及びオーディオ信号124を生成し得る。すなわち、セグメント(例えば、位置又は時間変化に基づいて定義されたシーン)の定義に基づいて、システム110は、ビデオフレーム、スクリプト(又は脚本)、ビデオのオーディオ、及び他のメタデータ情報の組み合わせを使用して、セグメント化を実行することができる。

【0063】

ビデオがセグメント化されると、分析モジュール114は、ビデオフレーム122、オーディオ信号124、及びテキスト情報126にそれぞれのAIモデルを適用することによって、セグメント146と関連付けられたコンテキスト情報170を生成することができる。コンテキスト情報170は、複数の入力モダリティのそれぞれの分類を融合させることによって取得されるマルチモーダル分類を含むことができる。続いて、注釈モジュール116は、セグメント146に融合又は統一分類を注釈付けすることができる。セグメント146と関連付けられた注釈210は、一組のキー212、214、216、218、及び220を含むことができる。これらのキーは、シーン142及びシーン142内の1つ以上のビートと関連付けられた意味コンテキスト情報を表すことができる。それぞれのキーは、フォーマット{キー:(値、強度)}であり得る。例えば、キー212は、値222及び強度224を含むことができる。

【0064】

キー212などのキーは、セグメント146と関連付けられた任意の意味コンテキスト情報を表すことができる。それ故に、キーは、新しいキーが加えられるにつれて拡張することができる拡張可能なリストとして表すことができる。典型的に、キー212は、セグメント146と関連付けられたオブジェクト、ジャンル、コンセプト、又はアクションを表すことができる。キー212はまた、評価、キャラクタ、及び位置を表すこともできる。キー212の値222は、キー212のタイプに依存することができ、また、セグメ

10

20

30

40

50

ント 1 4 6 の特徴であり得る。強度 2 2 4 は、所与の値（例えば、強度レベル又は信頼レベル）が存在する可能性を示す数字（例えば、0～1）であり得る。例えば、キー 2 1 2 がセグメント 1 4 6 のカテゴリ（又はジャンル）を表す場合、値 2 2 2 は、アクション、アドベンチャー、コメディ、クライム、ドラマ、ファンタジー、ホラー、ミュージカル、ミステリー、ロマンス、及び S F のうちの 1 つであり得る。一方で、キー 2 1 2 がセグメント 1 4 6 の評価を表す場合、値 2 2 2 は、バイオレンス、言語、ヌーディティ、及びホラーのうちの 1 つであり得る。

【0 0 6 5】

次いで、強度 2 2 4 は、カテゴリが特定のジャンルである可能性を示すことができる。したがって、値 2 2 2 が「アクション」である場合、強度 2 2 4 は、セグメント 1 4 6 がアクションジャンルである可能性を示すことができる。図 2 A に示すように、セグメント 1 4 6 がムービー X の戦場シーンに対応する場合、強度 2 2 4 は、値 2 2 2 が「アクション」である場合、1 に近くなり得る。次いで、セグメント 1 4 6 の例示的なキーは、{「ジャンル」：（「アクション」、0. 7 1）}、{「ジャンル」：（「コメディ」、0. 3 3）}、{「ジャンル」：（「親友」、0. 2）}、{「オブジェクト」：（「人」、0. 9）}、{「コンセプト」：（「戦争」、0. 9 5）}、{「アクション」：（「戦闘」、0. 9）}、及び {「評価」：（「バイオレンス」、0. 9）} を含み得るが、これらに限定されない。ジャンルなどのいくつかのキーは、互いに排他的でない場合がある。例えば、セグメント 1 4 6 は、1 つを超えるジャンルと関連付けることができる。

【0 0 6 6】

キー 2 1 2、2 1 4、2 1 6、2 1 8、及び 2 2 0 が意味カテゴリ（例えば、ジャンル）を表す場合、注釈モジュール 1 1 4 は、キーと関連付けられた意味カテゴリ（すなわち、値）の正規化された表現を生成することができる。注釈 2 1 0 をセグメント 1 4 6 に組み込むために、注釈モジュール 1 1 4 は、各キーと関連付けられた値の強度の正規化された表現を示すことができる。キー 2 1 2、2 1 4、2 1 6、2 1 8、及び 2 2 0 の値がカテゴリを表すことができるので、強度は、セグメント 1 4 6 がカテゴリに属する可能性に対応することができる。正規化は、1（又は 1 0 0）の値に基づくことができる。セグメント 1 4 6 が特定のカテゴリにある可能性は、注釈値によって 0～1 で表され、全てのカテゴリの注釈値の和は 1 であり得る。いくつかの実施形態では、セグメント 1 4 6 と関連付けられた注釈値の各々は、独特な表現で（例えば、色、フィルパターン、又は値で）表現し、注釈インジケータ 1 4 4 内に表すことができる。続いて、システム 1 1 0 は、注釈インジケータをビデオプロバイダに提供することができ、次いで、セグメント 1 4 6 を有する注釈インジケータ 1 4 4 を表示することができる。

【0 0 6 7】

注釈 2 1 0 は、視聴者のパーソナライズされた視聴オプションを容易にすることができる。システム 1 1 0 は、ユーザが、注釈 2 1 0 の一組のキー 2 1 2 の値を選択することを可能にする。視聴者は、特定のカテゴリに属するビデオ 1 4 0 のアイテムを自動的にスキップすることを選択し得る。強度 2 2 4 が閾値よりも大きい場合は、そのセグメント 1 4 6 がキー 2 2 2 に対応するカテゴリに属することを示し得る。視聴者がカテゴリをスキップすることを望む場合は、ビデオ 1 4 0 を再生するときに、セグメント 1 4 6 を自動的にスキップすることができる。一方で、視聴者はまた、そのカテゴリに属するセグメントを見ることも選択し得る。そのような状況下で、ビデオ 1 4 0 を再生するときに、カテゴリに属するセグメントのみを再生することができる。システム 1 1 0 は、セグメント 1 4 6 の開始及び終了タイムスタンプを維持することができる。その結果、セグメント 1 4 6 をスキップするために、セグメント 1 4 6 の開始タイムスタンプが検出されるたびに、ビデオ 1 4 0 を終了タイムスタンプまでスキップすることができる。同様に、セグメント 1 4 6 を示すために、カテゴリに属する以前のセグメントが再生を終了するたびに、ビデオ 1 4 0 は、セグメント 1 4 6 の開始タイムスタンプから再生を続ける。

【0 0 6 8】

図 2 B は、本出願の一実施形態による、ビデオセグメントと関連付けられた例示的な

ローカライズされた注釈を示す。ローカライズされた注釈は、シーンレベル注釈及びビートレベル注釈を含むことができる。これは、広告システムなどの異なるアプリケーションが注釈 250 を利用することができるような方法で、システム 110 が、ビデオセグメント 250 に注釈付けすることを可能にする。例えば、セグメント 250 の注釈は、セグメント 256 全体のセグメント注釈 256、セグメント 250 の第 1 のビートの第 1 のビート注釈 252、及びセグメント 250 の最後のビートの最後のビート注釈 254 を含むことができる。ビートレベルのキーは、セグメント 250 の対応するビートと関連付けられた意味コンテキスト情報を表すことができる。ここで、注釈 254 は、セグメント 250 の最後のビートと関連付けられた一組のキーを含むことができる。

【0069】

注釈 254 のそれぞれのキーは、フォーマット {キー: (値、強度)} であり得る。例えば、注釈 254 のキー 260 は、値 262 及び強度 264 を含むことができる。次いで、注釈 254 の例示的なキーは、{「オブジェクト」: (「人」、0.7)}、{「オブジェクト」: (「顔」、0.85)}、{「ジャンル」: (「コメディ」、0.85)}、{「コンセプト」: (「幸福」、0.95)}、及び {「アクション」: (「微笑み」、0.8)} を含み得るが、これらに限定されない。いくつかのキーは、互いに排他的でない場合がある。例えば、ビートは、1 つを超えるオブジェクトと関連付けることができる。注釈 254 は、アプリケーションをコンテキスト方式で操作するのを補助することができる。アプリケーションが広告配置に関連する場合、アプリケーションは、注釈 254 によって、セグメント 250 が楽しい雰囲気では終了したと判断することができ、セグメント 250 の後に配置される任意の広告は、幸福の話であると語るはずである。したがって、セグメントレベル及びビートレベル注釈を組み合わせたことは、セグメント 250 を効率的かつ正確にコンテキスト化することができる。

システムアーキテクチャ

【0070】

図 3A は、本出願の一実施形態による、分類ベースのマルチモーダル融合を使用して、ビデオセグメントのローカライズされたコンテキスト注釈を生成する例示的なプロセスを示す。システム 110 は、複数の入力モダリティを通じて同時推論から意味コンテキスト情報を推定することができる。操作中に、システム 110 は、入力ビデオ 300 を取得することができる。次いで、システム 110 は、シーン（又はショット）セグメント化技術 302 を使用してビデオ 300 を処理して、ビデオ 300 を、セグメント 320 などの一組のセグメント（例えば、ビデオ 300 の組成のシーン又はショット）に分離させることができる。ここで、セグメント 320 は、ビデオフレーム、オーディオ信号、及びテキスト情報を含み得る。システム 110 は、セグメント 320 を、ビデオフレーム 322 などの構成信号モダリティ、オーディオ信号 324、及び利用可能であれば、サブタイトル／字幕 336 などのテキスト情報 326 に更に解析することができる。システム 110 は、オーディオ／ビデオ（AV）コンバータ及びサブタイトル抽出器などのビデオ編集ツールを使用して、それぞれのセグメントの解析操作を実行し得る。

【0071】

代替的に、システム 110 は、ビデオ 300 を解析して、ビデオ 300 のビデオフレーム、オーディオ信号、及びテキスト情報を取得し得る。次いで、システム 110 は、ビデオ 300 のビデオフレームを、そのうちの 1 つがビデオフレーム 322 であり得るいくつかのビデオフレームセグメントにセグメント化し、対応するタイムスタンプを決定することができる。次いで、システム 110 は、抽出されたオーディオ及びテキスト情報をセグメント化して、対応するビデオフレームセグメントと整列させ、それにより、オーディオ信号 324 及びテキスト情報 326 を生成することができる。

【0072】

システム 110 は、ビデオ 300 のスクリプト 338 などの外部の入力モダリティを更に組み込むことができる。スクリプト 338 は、テキスト情報 326 の他の形態とみなすことができ、よって、サブタイトル／字幕 336 及びスクリプト 338 を含むことがで

10

20

30

40

50

きる。システム 110 は、ビデオフレーム 322 を処理することができ、これは、ビデオフレーム 322 内の画素データからシーンカテゴリ又はクラスを推論することができるディープビジュアルネットワーク 304 を含むことができる視覚認識技術を使用して処理される。ディープビジュアルネットワーク 304 は、畳み込みニューラルネットワーク (convolutional neural network、CNN) などのディープニューラルネットワーク (deep neural network、DNN) を含むことができる。更に、時間的力学を捕捉するために、ディープビジュアルネットワーク 304 は、CNN と併せてリカレントニューラルネットワーク (recurrent neural network、RNN) を含み得る。ディープビジュアルネットワーク 304 における CNN 及び RNN の組み合わせは、予めラベル付けした訓練データに基づいて、意味クラス、オブジェクト、及び／又はセグメント 320 内の人々を推論する (すなわち、対応するラベルを決定する) ように訓練することができる。検出されたクラス、オブジェクト、及び／又は人々は、セグメント 320 のビジュアル意味コンテキスト情報であり得る。

10

【0073】

システム 110 は、ディープスタックオートエンコーダ 306 を含むことができる別個のオーディオ処理技術を使用して、順次的又は並列的にオーディオ信号 324 を分析することができる。オートエンコーダ 306 は、オーディオ信号 324 の潜在的低次元埋め込みを学習することができる。オートエンコーダ 306 は、音楽及び特殊効果などのバックグラウンドの非言語信号 332、並びに人間の音声信号 334 をそれぞれ復元する、2 つのデコーダストリームが続くエンコーダを含み得る。言語的及び非言語的な信号の分離は、シーン特性を決定するために使用することができる。例えば、バックグラウンド音楽は、しばしば、ジャンル、ペース、テンション、などを表す。訓練中に、オートエンコーダ 306 は、分離させたバックグラウンド信号 332 及び音声信号 334 を再混合するための暗黙的モデルに基づいて、オーディオ信号 324 を再構築するように学習することができる。混合モデルは、メル周波数セプストラル係数などの適切な表現に基づいて操作することができる。

20

【0074】

続いて、システム 110 は、関連するシーン特性を推論する (すなわち、対応するラベルを決定する) ことができるオーディオバックグラウンド分類子 312 を使用して、バックグラウンド信号 332 を処理することができる。分類子 312 は、サポートベクターマシン又はランダムフォレストなどの機械学習技術と組み合わせて、スペクトログラムなどの入力ベースの特徴を含み得る。分類子 312 はまた、シーン特性を直接推論するために、エンドツーエンドのビデオ信号で訓練された DNN でもあり得る。分類子 312 はまた、ハイブリッドモデルを含むこともできる。ハイブリッドモデルは、DNN を使用して、判別可能な埋め込みを決定し得る。次いで、システム 110 は、差別的埋め込みを、ランダムフォレスト分類子などの非ディープラーナーに提供することができる。

30

【0075】

一方で、システム 110 は、長短期メモリ (Long-Short-Term Memory、LSTM) に基づく RNN などのメモリ構成要素を有する DNN を含むことができる音声認識技術 314 を使用して、音声信号 334 を処理することができる。音声認識技術 314 は、音声信号 334 からテキストを生成することができる。システム 110 は、音声認識技術 314 によって生成されたテキストに感情検出技術 316 を使用して、セグメント 320 の感情及び印象を推論することができる。更に、システム 110 は、音声認識技術 314 の代わりに、又はそれを強化するためにサブタイトル／字幕 336 を使用することができる。システム 110 は、単語／語句位置合わせ技術 318 を使用して、音声認識 314、サブタイトル／字幕 336、及びスクリプト 338 の組み合わせから、テキストを整列させ、強化することができる。次いで、システム 110 は、整列させたテキストをテキストベースのシーン分類子 330 に、入力として提供することができる。

40

【0076】

分類子 330 は、整列させたテキストから、セグメント 320 と関連付けられた心理

50

、感情、及びジャンルなどのコンテキスト情報を推論（すなわち、対応するラベルを決定）し得る、有意な埋め込みを抽出することができる。感情検出技術 3 1 6 及び／又は分類子 3 3 0 は、自然言語処理（NLP）に基づくことができる。次いで、システム 1 1 0 は、分類子スコア溶融技術 3 1 0 に基づいて、ビジュアル、オーディオ、及びテキスト分類子の各々から分類、予測、及び／又はスコアを融合させ、セグメント 3 2 0 と関連付けられた複数の入力モダリティから同時推論されたコンテキスト情報を生成することができる。分類子スコア溶融技術 3 1 0 は、多数決分類及び／又は重み付き分類を使用して、セグメント 3 2 0 と関連付けられた分類を決定し得る。システム 1 1 0 の注釈モジュール 1 1 6 は、コンテキスト情報を取得し、セグメント 3 2 0 の注釈 3 4 0 を生成することができる。注釈 3 4 0 は、コンテキスト情報を表す一組のキーを含むことができる。システム 1 1 0 は、入力ビデオ 3 0 0 のそれぞれのセグメントに対してこのプロセスを繰り返すことができる。

10

【0077】

ここで、分類子スコア溶融技術 3 1 0 は、ビジュアル、オーディオ、及びテキスト分類子から生成される分類から取得される分類に適用される。したがって、分類子スコア融合技術 3 1 0 は、「後期融合」戦略とみなすことができる。個々の入力モダリティに基づいて分類に対して後期融合を実行する代わりに、システム 1 1 0 はまた、入力モダリティから前期又は中期レベルの埋め込みを融合させ、統一マルチモーダル分類子を融合マルチモーダル埋め込みに適用し得る。図 3 B は、本出願の一実施形態による、埋め込みベースのマルチモーダル融合を使用して、ビデオセグメントのローカライズされたコンテキスト注釈を生成する例示的なプロセスを示す。ここで、システム 1 1 0 は、分類ベースのマルチモーダル融合の代わりに、埋め込みベースのマルチモーダル融合を使用することができる。

20

【0078】

埋め込みベースのマルチモーダル融合は、セグメント 3 2 0 の入力モダリティの各々に対して導出されたベクトル埋め込みを取得し、それらをマルチモーダル埋め込みに融合させる。例えば、システム 1 1 0 は、ディープビジュアルネットワーク 3 0 4 からのビジュアル埋め込み 3 5 2、オートエンコーダ 3 0 6 からのバックグラウンド信号埋め込み 3 5 4、及びテキストベースの分類子 3 3 0（例えば、word2vec NLP モデル）からの単語埋め込み 3 5 6 を取得することができる。いくつかの実施形態では、システム 1 1 0 は、DNN を使用して、対応する分類子からの埋め込みを自動的に学習することができる。特定の入力モードと関連付けられた埋め込みが、対応する分類子の中間状態からユーザによって取得することができることに留意されたい。

30

【0079】

次いで、システム 1 1 0 は、埋め込み融合／連結技術 3 6 0 をこれらの埋め込みに適用して、マルチモーダル埋め込み 3 5 8 を生成することができる。システム 1 1 0 は、ディープビジュアルネットワーク 3 0 4、オートエンコーダ 3 0 6、及びテキストベースの分類子 3 3 0 から特徴埋め込みを取得し、この埋め込みをマルチモーダル埋め込み 3 5 8 を表す連結層において融合させることができる。次いで、システム 1 1 0 は、マルチモーダル埋め込み 3 5 8 をマルチモーダルセグメント分類子 3 6 2 に、入力として提供することができる。分類子 3 6 2 は、マルチモーダル埋め込み 3 5 8 に基づいて、セグメント 3 2 0 に重要な特性を推論することができる。そうするために、次いで、システム 1 1 0 は、連結層を分類子 3 6 2 に組み込むことができ、これを次に、分類ラベルを出力層に提供することができる。

40

【0080】

いくつかの実施形態では、システム 1 1 0 は、埋め込みベース及び／又は分類ベースのマルチモーダル融合の変形例を使用し得る。例えば、オーディオ及びテキストと関連付けられた埋め込み 3 5 4 及び 3 5 6 は、それぞれ、後のステージ内の適切な層のディープビジュアル分類子 3 0 4 に挿入することができる。次いで、埋め込み 3 5 4 及び 3 5 6 を組み込むことによって、分類子 3 0 4 を、セグメント 3 2 0 と関連付けられたマルチモー

50

ダルコンテキスト情報を推論するように訓練することができる。次いで、システム 110 の注釈モジュール 116 は、コンテキスト情報を取得し、セグメント 320 の注釈 340 を生成することができる。注釈 340 は、コンテキスト情報を表す一組のキーを含むことができる。システム 110 は、入力ビデオ 300 のそれぞれのセグメントに対してこのプロセスを繰り返すことができる。

広告システム

【0081】

システム 110 によって生成される注釈は、広告システムなどの他のアプリケーションに使用することができる。広告システムは、システム 110 から取得される注釈に基づいて、ビデオのコンテキスト的に関連する広告を配置することができる。図 4 は、本出願の一実施形態による、ローカライズされたコンテキスト情報に基づく、ビデオの例示的な広告配置プロセスを示す。セグメント化モジュール 112 は、ビデオ 420 を取得することができ、ビデオ 420 をいくつかのセグメント 422、424、426、及び 428 にセグメント化することができる。分析モジュール 114 は、セグメントの複数の入力モダリティに基づいて、それぞれのセグメントと関連付けられたコンテキスト情報を推論することができる。

【0082】

注釈モジュール 116 は、コンテキスト情報を使用して、セグメント 422、424、426、及び 428 の注釈 442、444、446、及び 448 をそれぞれ生成することができる。注釈モジュール 116 はまた、セグメント 424 の最後のビートの注釈 552 及びセグメント 426 の最初のビートの注釈 554 などの、それぞれのセグメントの 1 つ以上のビートのビートレベル注釈を生成することもできる。次いで、システム 110 の広告モジュール 118 は、ビデオ 420 内の ISA 432、434、及び 436 を、潜在的広告スポットとして識別することができる。このようにして、ビデオ 420 の ISA の広告の配置は、各セグメントがシーンを含むことができるので、ビデオ 420 内のシーンを中断し得ない。

【0083】

管理者は、広告を示すためにビデオ 420 を停止させることができる回数を構成し得る。この数に基づいて、広告モジュール 118 は、ビデオ 420 の広告を配置するための一組の ISA を推奨し得る。広告モジュール 118 は、1 つ以上のポリシーを注釈に適用して、広告を配置するために使用することができる一組の ISA を選択し得る。そのようなポリシーは、マネタイゼーションを最大にすること、ハラハラする展開を回避する又は使用すること、シーンの断片化を回避すること、及び心地よい先行セグメントを含むことができるが、これらに限定されない。ビデオ 420 を停止することができる回数が 1 に構成されると仮定する。次いで、広告モジュール 118 は、広告を配置するために、ISA 432、434、及び 436 のうちの 1 つを選択することができる。例えば、広告モジュール 118 は、ユーザが広告の物語の影響を受けやすくなり得る ISA を選択し得る。広告モジュール 118 は、（例えば、意味カテゴリ及びそれらの対応する信頼値に適用された重み付けスコアに基づいて）注釈を調査し、感情的な情報を評価し、また、感受性を決定することができる。

【0084】

広告モジュール 118 が ISA 434 を選択した場合、広告モジュール 118 は、注釈情報 410 を広告システム 400（例えば、広告交換及び／又はプラットフォーム）に提供することができる。注釈情報 410 は、ISA 434 の先行セグメント 424 及び後続セグメント 426 の注釈 444 及び 446 をそれぞれ含むことができる。注釈情報 410 はまた、セグメント 424 内の最後のビートの注釈 552 及びセグメント 426 内の最初のビートの注釈 554 を含むこともできる。広告システム 400 は、利用可能な広告並びにそれらのそれぞれのコンテンツ及び物語（例えば、各広告の一組のキーワードとして表現される）をデータベース 402 内に維持することができる。

【0085】

注釈情報 410 を取得すると、広告システム 400 は、注釈情報 410 内の注釈と、ISA 434 内に配置することができる対応する広告とを動的にマッチさせることができる。マッチング操作を実行するために、広告システム 400 は、注釈内のキーから値及び強度を取得し、閾値よりも大きい対応する強度を有する一組の値を選択することができる。次いで、広告システム 400 は、コンテンツを説明するキーワード及び利用可能な広告と、選択された値の 1 つ以上とをマッチさせ、最良のマッチを有する広告を選ぶことができる。マッチング操作は、データベース 402 で実行される 1 つ以上のデータベースクエリに基づくことができる。

【0086】

代替的に、広告を所有する企業又は企業の代理人は、広告システム 400 から、事業に関連するタグを購入することができる。例えば、企業が保険会社である場合、企業は、事故、火事、爆発などに関連するタグを購入することができる。セグメント 424 及び／又はセグメント 426 が、カーチェイス及び／又は車両の損傷を含むアクションシーケンスに対応する場合、注釈情報 410 の注釈は、アクションを表す。注釈内の対応する値を広告のタグにマッチさせることができるので、広告システム 400 は、広告が ISA 434 内に配置され得ると判定することができる。このようにして、注釈は、ビデオ 420 内のコンテキスト的に関連する広告の配置を容易にすることができる。

【0087】

典型的に、ビデオ 420 が映画などの一般的なビデオである場合、ビデオ 420 のプロダクトプレイスメントは、広告の有効な形態であり得る。しかしながら、そのようなプロダクトプレイスメントは、高価であり得る。ビデオ 420 内のコンテキスト広告は、費用効果的な変形例になり得る。視聴者がセグメント 424 で主人公がハイウェイ上で高級車を運転するのを見る場合、コンテキスト広告は、広告システム 400 が高級車の広告を ISA 434 内のセグメント 424 の直後に示すことを可能にする。コンテキスト手法はまた、広告オーバーレイでも使用することができる。更に、いくつかのビデオプロバイダは、サブスクリプションを必要とするので、ビデオ提供者はまた、視聴者情報（例えば、性別、人種、年齢、所得階層、など）も知り得る。視聴者情報及びコンテキスト情報を組み合わせることで、ビデオ 420 内のコンテキスト広告の効果を更に向上させることができる。

操作

【0088】

図 5A は、本出願の一実施形態による、分類ベースのマルチモーダル融合を使用して、ビデオセグメントの統一分類を生成する、ビデオ注釈システムの方法 500 を示すフローチャートを示す。操作中に、システムは、利用可能である場合に入力ビデオ及び関連付けられたスクリプトを取得し、セグメント化技術に基づいて、入力ビデオのビデオセグメントを決定することができる（操作 502）。次いで、システムは、それぞれのビデオセグメントを解析して、ビデオフレーム及びオーディオ信号を、並びに利用可能である場合にビデオセグメントのサブタイトル／字幕を取得することができる（操作 504）。続いて、システムは、オーディオエンコーダを適用して、音声及びバックグラウンド信号をオーディオ信号から分離することができる（操作 506）。

【0089】

システムは、オーディオ分類に対して、オーディオ分類子をバックグラウンド信号に適用し、感情分類に対して、感情分類子を音声信号に適用することができる（操作 508）。システムは、ビジュアル分類に対して、ディープビジュアルネットワークをビデオフレームに更に適用することができる（操作 510）。システムはまた、音声認識を使用して音声信号から音声テキストを決定し、サブタイトル／字幕及びスクリプトを音声テキストと整列させることもできる（操作 512）。次いで、システムは、テキスト分類に対して、テキストベースの分類子（例えば、NLP ベースの分類子）を、整列させたテキストに適用することができる（操作 514）。続いて、システムは、ビデオセグメントに対して、マルチモーダル分類から統一分類を決定することができる（操作 538）。

【0090】

図5Bは、本出願の一実施形態による、埋め込みベースのマルチモーダル融合を使用して、ビデオセグメントの統一分類を生成する、ビデオ注釈システムの方法530を示すフローチャートを示す。操作中に、システムは、利用可能である場合に入力ビデオ及び関連付けられたスクリプトを取得し、セグメント化技術に基づいて、入力ビデオのビデオセグメントを決定することができる（操作532）。次いで、システムは、それぞれのビデオセグメントを解析して、ビデオフレーム及びオーディオ信号を、並びに利用可能である場合にビデオセグメントのサブタイトル／字幕を取得することができる（操作534）。続いて、システムは、オーディオエンコーダを適用して、音声及びバックグラウンド信号をオーディオ信号から分離することができる（操作536）。

10

【0091】

システムは、オーディオ特徴埋め込みを判定するために、オーディオ分類子をバックグラウンド信号に適用することができる（操作538）。システムは、ビジュアル特徴埋め込みを判定するために、ディープビジュアルネットワークをビデオフレームに更に適用することができる（操作540）。システムはまた、音声認識を使用して音声信号から音声テキストを決定し、サブタイトル／字幕及びスクリプトを音声テキストと整列させることもできる（操作542）。次いで、システムは、テキスト特徴埋め込みを決定するために、テキストベースの分類子を、整列させたテキストに適用することができる（操作544）。続いて、システムは、特徴埋め込みを組み合わせ、統一特徴埋め込みを生成し（操作546）、マルチモーダルセグメント分類子を統一特徴埋め込みに適用して、ビデオセグメントクラスを決定することができる（操作548）。

20

【0092】

図5Cは、本出願の一実施形態による、ビデオセグメントの統一分類に基づいて注釈を生成する、ビデオ注釈システムの方法550を示すフローチャートを示す。操作中に、システムは、ビデオセグメントの統一分類を取得し（操作552）、ビデオセグメントの統一分類に基づいて一組のキーを決定することができる（操作554）。ユーザは、ビデオセグメントに対して、一組のキーを選択し、構成することができる。次いで、システムは、統一分類のコンテキスト情報からそれぞれのキーを取得し、キーの値及び強度を決定することができる（操作556）。続いて、システムは、一組のキーを有するビデオセグメントの注釈を生成することができる（操作558）。

30

【0093】

図6は、本出願の一実施形態による、広告をビデオに配置するための注釈情報を提供する、ビデオ注釈システム600の方法を示すフローチャートを示す。操作中に、システムは、入力ビデオのISAを決定することができ（操作602）、ビデオの広告配置の数を決定することができる（操作604）。次いで、システムは、1つ以上の選択ポリシーに基づいて、広告配置の数に対応する一組のISAを選択することができる（操作606）。続いて、システムは、ビデオセグメントに対応する注釈、及び選択された一組のISAと関連付けられたビデオセグメントのビートを取得することができる（操作608）。システムは、注釈を広告システムに提供することができる（操作610）。

40

例示的なコンピュータシステム及び装置

【0094】

図7は、本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈をサポートする、ビデオ注釈システムを容易にする例示的なコンピュータシステムを示す。コンピュータシステム700は、プロセッサ702、メモリデバイス704、及び記憶デバイス708を含む。メモリデバイス704は、揮発性メモリデバイス（例えば、デュアルインラインメモリモジュール（dual in-line memory module、DIMM））を含むことができる。更に、コンピュータシステム700は、ディスプレイデバイス710、キーボード712、及びポインティングデバイス714に連結することができる。記憶デバイス708は、オペレーティングシステム716、ビデオ注釈システム718、及びデータ736を記憶することができる。ビデオ注釈システム718は、システム110の操作を

50

組み込むことができる。

【0095】

ビデオ注釈システム718は、コンピュータシステム700によって実行されると、コンピュータシステム700に、本開示に記載される方法及び／又はプロセスを実行させることができる命令を含むことができる。具体的には、ビデオ注釈システム718は、ビデオをビデオセグメントにセグメント化ユニットに基づいてセグメント化するための命令を含むことができる（セグメント化論理ブロック720）。ビデオ注釈システム718はまた、複数の入力モダリティを取得するために、それぞれのビデオセグメントを解析するための命令を含むこともできる（セグメント化論理ブロック720）。更に、ビデオ注釈システム718は、それぞれのビデオセグメントのビデオフレームの意味コンテキスト分類を取得するための命令を含む（ビデオ論理ブロック724）。 10

【0096】

ビデオ注釈システム718はまた、それぞれのビデオセグメントの音声チャンネルからバックグラウンド及び音声信号を分離するための命令を含むこともできる（音声論理ブロック724）。ビデオ注釈システム718は、音声信号からバックグラウンド信号及び音声テキストの意味コンテキスト分類を取得するための命令を更に含むことができる（音声論理ブロック724）。更に、ビデオ注釈システム718は、音声テキストを対応するビデオセグメントのテキスト情報と整列させるための命令を含むことができる（テキスト論理ブロック726）。ビデオ注釈システム718はまた、整列させたテキストの意味コンテキスト分類を取得するための命令を含むこともできる（テキスト論理ブロック726） 20

【0097】

ビデオ注釈システム718は、マルチモーダル分類を融合させ、統一分類を決定するための命令を含むことができる（融合論理ブロック728）。ビデオ注釈システム718はまた、マルチモーダル分類からの埋め込みを融合させて、マルチモーダル埋め込みを決定し、マルチモーダル埋め込みから統一分類を決定するための命令を含むこともできる（融合論理ブロック728）。更に、ビデオ注釈システム718は、統一分類からそれぞれのビデオセグメントの注釈を生成するための命令を含むことができる（注釈論理ブロック730）。 30

【0098】

いくつかの実施形態では、ビデオ注釈システム718は、ビデオ内のISAを決定し、ISAと関連付けられた注釈情報を広告システムに提供するための命令を含むことができる（広告論理ブロック732）。ビデオ注釈システム718は、メッセージ（例えば、パケット及びフレーム）を送信及び受信するための命令を更に含み得る（通信論理ブロック734）。データ736は、システム110の操作を容易にすることができる任意のデータを含むことができる。データ736は、異なる分類子の訓練データ、1つ以上のビデオのセグメントのマルチモーダル情報、セグメントの分類及び埋め込み情報、並びにビデオのISA情報のうちの1つ以上を含み得る。 30

【0099】

図8は、本出願の一実施形態による、ビデオのローカライズされたコンテキスト注釈を容易にする、例示的な装置を示す。ビデオ注釈装置800は、有線、無線、量子光、又は電気通信チャンネルを介して互いに通信し得る複数のユニット又は装置を含むことができる。装置800は、1つ以上の集積回路を使用して実現され得、図8に示されているものよりも少ない又は多いユニット又は装置を含み得る。更に、装置800は、コンピュータシステムに統合され得るか、又は他のコンピュータシステム及び／又はデバイスと通信することができる別個のデバイスとして実現され得る。具体的には、装置800は、図7のコンピュータシステム700のモジュール720～734に類似する機能又は操作を実行するユニット802～816を備えることができ、これらのユニットは、セグメント化ユニット802、ビデオユニット804、オーディオユニット806、テキストユニット808、融合ユニット810、注釈ユニット812、広告ユニット814、及び通信ユニッ 40 50

ト 8 1 6 を含む。

【 0 1 0 0 】

「発明を実施するための形態」に記載されるデータ構造及びコードは、典型的には、コンピュータ可読記憶媒体に記憶され、コンピュータ可読記憶媒体は、コンピュータシステムが使用するためのコード及び／又はデータを記憶することができる任意のデバイス又は媒体であり得る。コンピュータ可読記憶媒体としては、揮発性メモリ、不揮発性メモリ、ディスク、磁気テープ、C D (compact disc、コンパクトディスク)、D V D (digital versatile disc、デジタル多用途ディスク、又はdigital video disc、デジタルビデオディスク)などの磁気及び光学ストレージデバイス、又は現在知られている若しくは今後開発されるコンピュータ可読メディア媒体を記憶できるその他の媒体が挙げられるが、これらに限定されない。

10

【 0 1 0 1 】

「発明を実施するための形態」の節に記載される方法及びプロセスは、上に論じられるようなコンピュータ可読記憶媒体内に記憶され得るコード及び／又はデータとして具体化され得る。コンピュータシステムが、コンピュータ可読記憶媒体上に記憶されたコード及び／又はデータを読み取って実行すると、コンピュータシステムは、データ構造及びコードとして具体化され、コンピュータ可読記憶媒体内に記憶された方法及び処理を実行する。

【 0 1 0 2 】

更に、上述の方法及びプロセスは、ハードウェアモジュールに含めることができる。例えば、ハードウェアモジュールとしては、特定用途向け集積回路(application-specific integrated circuit、A S I C)チップ、フィールドプログラマブルゲートアレイ(field-programmable gate array、F P G A)、及び現在知られている又は今後開発される他のプログラム可能論理デバイスを含むことができるが、これらに限定されない。ハードウェアモジュールが起動されると、ハードウェアモジュールは、ハードウェアモジュール内に含まれる方法及びプロセスを実行する。

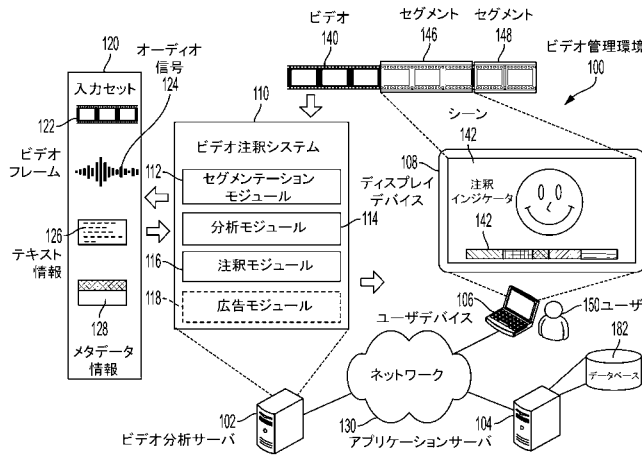
20

【 0 1 0 3 】

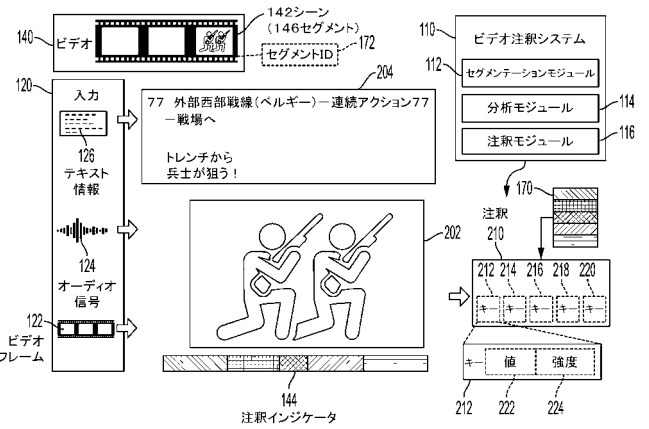
本明細書に記載される前述の実施形態は、例示及び説明のみを目的として提示されている。これらは、網羅的であること、又は本明細書に記載される実施形態を開示される形態に限定することを意図するものではない。したがって、多くの修正及び変形が、当業者には明らかであろう。加えて、上記の開示は、本明細書に記載される実施形態を限定することを意図するものではない。本明細書に記載される実施形態の範囲は、添付の特許請求の範囲によって定義される。

30

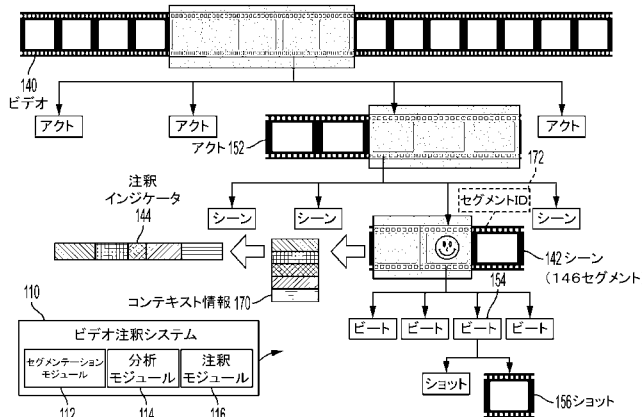
【図 1 A】



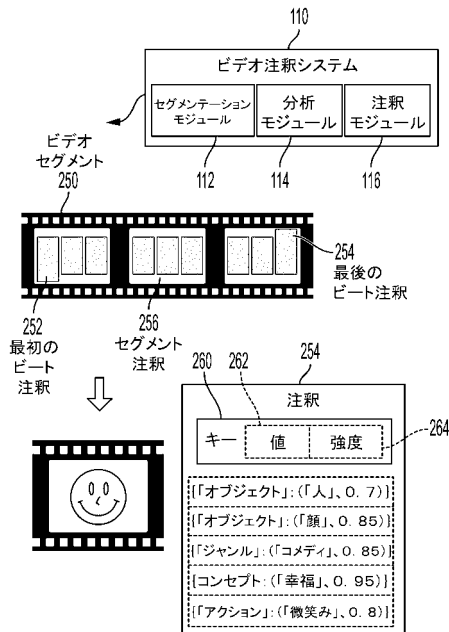
【図 2 A】



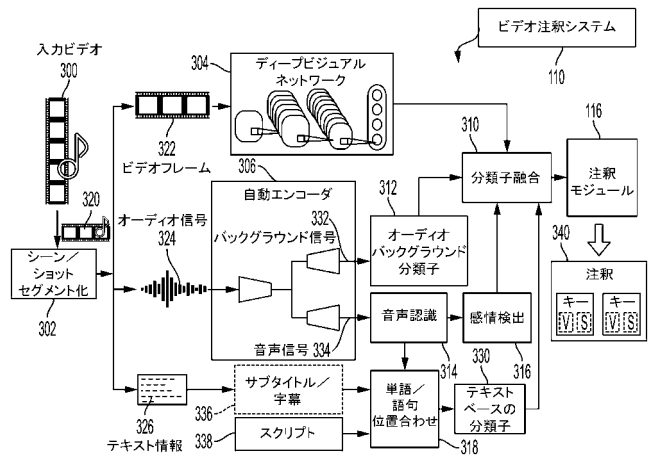
【図 1 B】



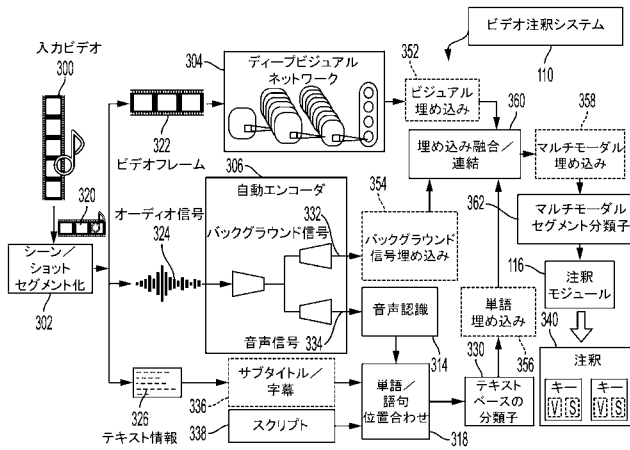
【図 2 B】



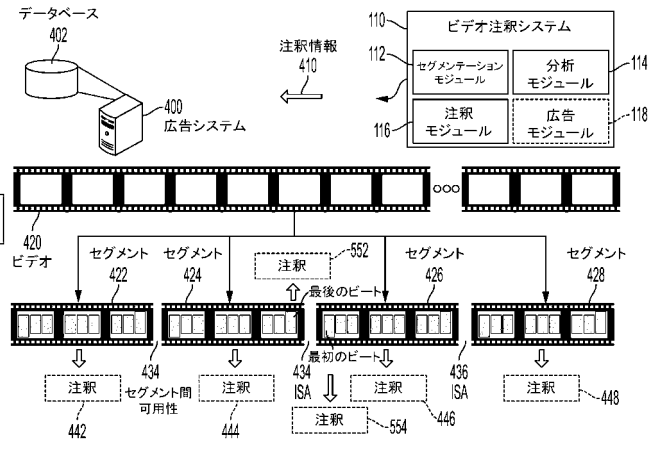
【図 3 A】



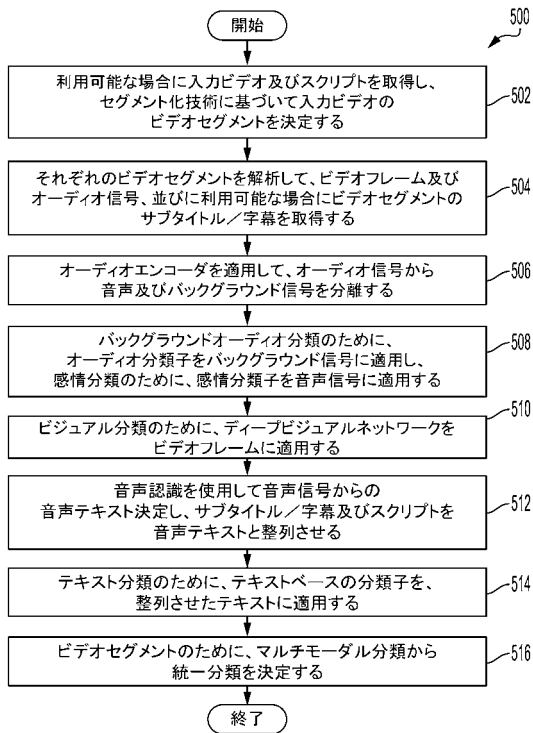
【図 3 B】



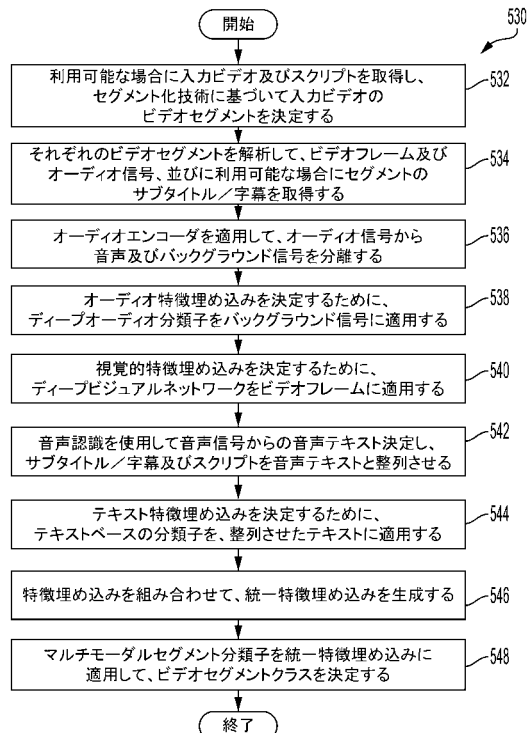
【図 4】



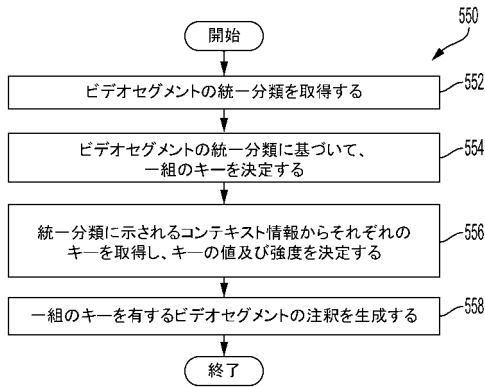
【図 5 A】



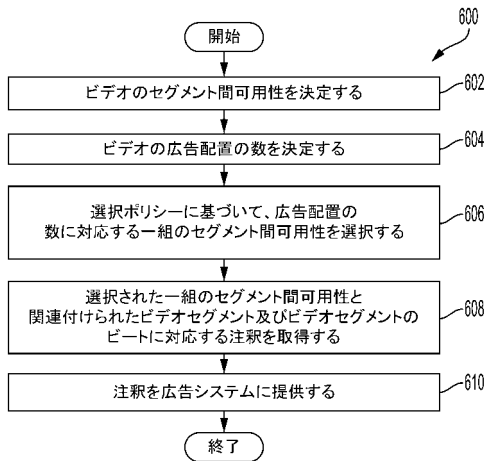
【図 5 B】



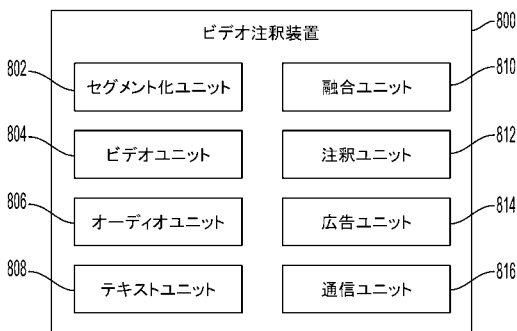
【図 5 C】



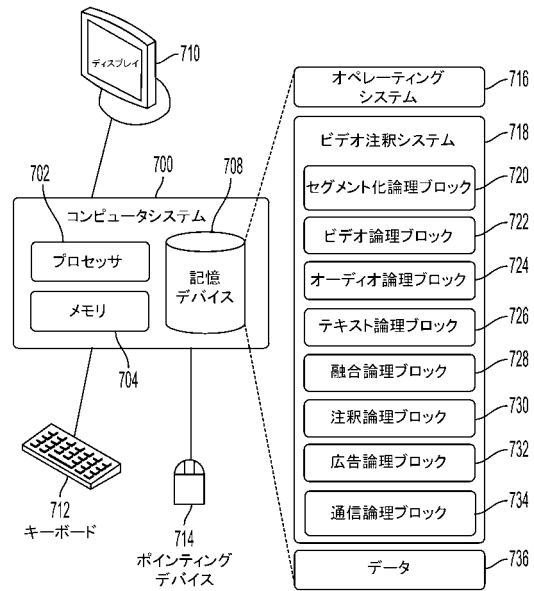
【図 6】



【図 8】



【図 7】



フロントページの続き

(74)代理人 100109335

弁理士 上杉 浩

(74)代理人 100120525

弁理士 近藤 直樹

(74)代理人 100139712

弁理士 那須 威夫

(74)代理人 100158551

弁理士 山崎 貴明

(72)発明者 カルナカラン・シュレシュクマル

アメリカ合衆国 カリフォルニア州 90025, ロサンゼルス, マルコムアヴェニュー 192
1, アpartment202

(72)発明者 ラジャ・バラ

アメリカ合衆国 ニューヨーク州 14534 ピッツフォード, ウッドグリーンドライブ 65

Fターム(参考) 5B175 FA03 FB02

5C164 MA06S MB44S PA44 SB01S SB04P SB06P SB31S